Lexis
Journal in English Lexicology

16 | 2020
Diachronic Lexical Semantics

Lexical Emergence on Reddit: An Analysis of Lexical Change on the "Front Page of the Internet"

Hanna Mahler

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/lexis/4917
DOI: 10.4000/lexis.4917
ISSN: 1951-6215

Publisher
Université Jean Moulin - Lyon 3

Electronic reference
Hanna Mahler, "Lexical Emergence on Reddit: An Analysis of Lexical Change on the "Front Page of the Internet"", *Lexis* [Online], 16 | 2020, Online since 17 December 2020, connection on 21 January 2021.
URL: http://journals.openedition.org/lexis/4917 ; DOI: https://doi.org/10.4000/lexis.4917

This text was automatically generated on 21 January 2021.

Lexis is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by OpenEdition

Lexical Emergence on Reddit: An Analysis of Lexical Change on the “Front Page of the Internet”

Hanna Mahler

Introduction

1 Current advancements in the availability and size of corpora have had considerable impact on linguistic research into historical semantics (Gries [2012], Sagi *et al.* [2012: 61]). Especially corpora of computer-mediated communication contain large amounts of data that make it possible to answer questions that could previously not have been asked (Grieve *et al.* [2017: 102]). Concerning diachronic semantic change, previous studies usually focus on decades or centuries (e.g. Geeraerts *et al.* [2012], Smith [2016]); with the new corpora, language change can be observed on a monthly, weekly, or even daily basis.

2 Based on these new possibilities, Grieve *et al.* [2017] draw attention to the lack of research on “lexical emergence”, which they define as “the process through which new word forms spread across a population of speakers” [2017: 102]. Currently available computational methods for identifying neologisms (e.g. Kerremans & Prokić [2018]) do not allow researchers to zoom in on this initial phase. Grieve *et al.* [2017] therefore present a methodology for finding emerging lexemes resulting from onomasiological change and apply it to a corpus of Twitter data. While their results are informative, the question remains how effective their methodology is when applied to different social media platforms, and whether the resulting word forms would be similar. The present study thus has a twofold goal: to test their methodology in a different context, the platform Reddit, and to compare the results. To be precise, the following two research questions will be addressed:
(i) Are the characteristics of emerging lexemes on Reddit similar to the characteristics of emerging lexemes on Twitter as identified by Grieve *et al.* [2017]?

(ii) How applicable is the methodology outlined by Grieve *et al.* [2017] for the study of lexical emergence on a different online platform, Reddit?

3 The present study is therefore located within the tradition of investigating onomasiological change from a usage-based perspective [Geeraerts 2006: 38], as well as recent approaches using large-scale corpora and quantitative methods to study lexical semantics and lexical change (see Allan & Robinson [2012], Geeraerts [2009: 233-235]). The study also widens the perspective of current research on computer-mediated communication to also consider online platforms other than Twitter, which is often the sole source of information. Furthermore, the trial and refinement of the methodology for discovering emerging lexemes holds valuable insights for scholars looking to apply this procedure in the future.

1. Theoretical background

4 The following section provides an overview of prior research on lexical change online and on the platform Reddit. Self-evidently, not all papers that contributed to the discussion can be rendered adequately within this short summary.

1.1. Lexical change in the online environment

5 Some terminological clarifications are in order at the start. This paper follows the definition of emerging lexemes by Grieve *et al.* [2017: 101] as new word forms that “spread across a population of speakers for the first time”. They can therefore be located at the initial stage of institutionalisation (Brinton & Traugott [2005: 45], Fischer [1998: 15]), also referred to as conventionalisation (Bakken [2006: 107]). While there appear to be diverging understandings of the term ‘neologism’ in the literature (as the outcome of institutionalisation [Brinton & Traugott 2005: 45] or the input to institutionalisation [Fischer 1998: 7]), the present study will conceptualize neologisms as a broader cover term comprising emerging lexemes.

6 The remainder of this section discusses important studies focusing on lexical change within computer-mediated communication. The main paper relevant for the present study is Grieve *et al.* [2017], who analyse ongoing lexical emergence in an 8.9 million words corpus of American Twitter data. From their data, they extract a total of 29 emerging lexemes that feature a high correlation coefficient (of the frequency of occurrence and the date of creation) over the whole year and that start off with a low overall frequency. They analyse these word forms concerning their parts of speech, word-formation process, time of origin, and competition with other lexemes. Their results suggest that emerging lexemes may be attested a long time before they increase in frequency and that their eventual spread follows the s-shaped curve (Blythe & Croft [2012], Nevalainen & Raumolin-Brunberg [2003: 53-55]) known from other linguistic phenomena (Grieve *et al.* [2017: 123-124]). Overall, they make the case for using large-scale web-based corpora for the analysis of lexical change, which the study at hand complies with. Grieve [2018] further analyses the “survival chances” of the lexemes identified by Grieve *et al.* [2017] over a longer time period and investigates which characteristics might contribute to their firm establishment. A similar attempt is made by Stewart & Eisenstein [2018], who try to predict word adoption on Reddit by early dissemination, and by Cole *et al.* [2017], who relate word adaptation to community size

on Reddit. Even though this aspect would have also been interesting, the question of how this process plays out on Reddit has to be postponed to future research.

7 A similar approach to lexical change in online communication is presented in Sang [2016]; he compares two methods of identifying neologisms and archaisms by applying them to two different corpora (Dutch magazines and Dutch tweets). In the first method, Sang compares the initial and final relative frequency of the lexemes in question, whereas in the second method, he calculates a correlation coefficient of the relative frequencies over different time periods (Sang [2016: 3-4]). He argues that both approaches can be used in tandem, since they produce different sets of rising and falling words. The neologisms he identifies are mainly English loan words into Dutch or results of spelling reforms (Sang [2016: 8]). Instead of comparing two methodologies, the present study focuses on applying and enhancing the second approach, which mostly overlaps with the description by Grieve *et al.* [2017].

8 Some studies also approach the issue of lexical emergence from a different perspective. Del Tredici & Fernández [2018], for example, investigate linguistic innovations on Reddit from a sociolinguistic viewpoint, based on Milroy’s network theory. Their findings, covering a number of topical sub-fora known as “subreddits”, suggest that linguistic ‘innovators’ are characterized as central members of each subreddit with weak-tie connections to other users, whereas ‘adaptors’, who are essential for the spread of a new item, have strong-tie connections to their specific sub-group (Del Tredici & Fernández [2018: 1]). Eisenstein *et al.* [2014], on the other hand, provide an analysis of lexical change on social media from the perspective of language change. Their corpus of American Twitter data reveals that innovations spread geographically from bigger to smaller cities within the United States, but that demographic similarity, especially ethnicity, also plays a substantial role in the dissemination of new lexical items (Eisenstein *et al.* [2014: 10]). These studies remind other researchers to keep extralinguistic factors, such as race, social network, or spatial proximity, in mind when investigating neologisms in computer-mediated communication (if this information is available).

9 This short overview shows how lexical emergence, its sociolinguistic and geographic aspects, have been investigated, primarily on Twitter. However, it remains unclear whether the same principles hold true for other online platforms. The present study fills this gap in knowledge by applying Grieve’s methodology to the online forum Reddit, which is described in more detail within the next section.

1.2. Reddit

10 The internet platform Reddit was founded in 2005 by Alexis Ohanian and Steve Huffmann. It is currently the fifth most visited website in the United States and consists of over 130,000 active communities [Reddit Inc. 2020]. The site is a “social news aggregation, web content rating, and discussion website” [Medvedev *et al.* 2017: 184] that can be viewed and joined for free. Content created by the users, also known as “redditors”, can be voted up or down by fellow users, while a reward system credits popular posts and comments with “karma”. On both Reddit and Twitter communication takes places asynchronously and both platforms have an upper character limit (40,000 characters for Reddit, 140 for Twitter in 2013). Furthermore, both allow for textual as well as visual modes of communication.

11 But there are also several important differences. Reddit is a highly anonymous platform, which makes it difficult to consider sociolinguistic metadata, whereas Twitter is more person-orientated. In addition, Reddit has a more differentiated internal structure with a plethora of subreddits covering a variety of topics. The participant characteristics as well as the tone and topic of the conversations vary heavily between these subreddits, which can be interpreted as individual communities of practice (Del Tredici & Fernández [2017: 3]). The illustration below (Figure 1) provides an example of a Reddit post followed by comments, taken from the subreddit “r/linguistics”:

Figure 1. Illustration of Reddit comment structure, taken from r/linguistics

12 Interested readers are also referred to the “Pushshift” website (Baumgartner [2020]), which contains frequently updated statistics on Reddit contributions and activity, as well as an overview of the most popular subreddits and contributors.

2. Methodology

13 The following sub-sections provide an in-depth description of methodological steps employed for gathering the results presented in Section 3. This includes notes on the corpus used, the processing of the data, as well as ethical considerations. This detailed account is necessary to enhance accountability and reproducibility, a goal every linguistic study should aim for (see Weller & Kinder-Kurlanda [2016: 168]).

2.1. The Pushshift Reddit Dataset

14 The data used in the study at hand is part of the Pushshift Reddit Dataset (Baumgartner *et al.* [2020]). The collection contains “all submissions and comments posted on Reddit between June 2005 and April 2019” [Baumgartner *et al.* 2020: 832]. The data is provided in JSON format. Each data point contains the raw text, as well as metadata consisting of the comment id, username, date of publication, status of the author, date of retrieval,

subreddit id, and whether the comment was edited, archived, or classified as "controversial". The uniqueness of this data set has been noted by several researchers, for example Weller & Kinder-Kurlanda [2016: 168], who discuss several approaches of sharing social media data.

15 While the large size of this data set is an asset, there are also some disadvantages that need to be taken into account. First of all, even though Reddit is an English platform, there are some contributions in other languages, which might lead to foreign language items interfering in the analysis. Furthermore, one cannot be sure whether the users are native speakers of English or not – deviant spellings or word usages might therefore not be innovations, but simply learner errors. Medvedev *et al.* [2018: 4] justly point out some further problems with the data set in question: considerable amounts of comments and posts appear to be missing in several years. However, they conclude that "[t]he risks of mis-sampled data are obvious, but in large scale studies they may be safely disregarded due to their smallness" [Medvedev *et al.* 2018: 4]. This is taken to be the case for the present study as well.

2.2. Data processing

16 The Pushshift Reddit Dataset data, which is sorted into months, was downloaded using the programme "µtorrent" (BitTorrent Inc. [2018]) and then unpacked using the "7-Zip" software (Pavlov [2018]). As a next step, all metadata was removed from the comments to yield monthly files with the raw text only. Afterwards, the "AntConc" programme (Anthony [2018]) was employed to create wordlists of each monthly sub-corpus. This proved to be a challenge for the application since the monthly data sets were up to 659 MB large. All the remaining steps were conducted with the help of R, a "language and environment for statistical computing and graphics" [R Core Team 2020]. Since the analysis by Grieve *et al.* [2017] is based on data from 2013 and 2014, the year 2013 was chosen as the temporal frame for all subsequent steps.

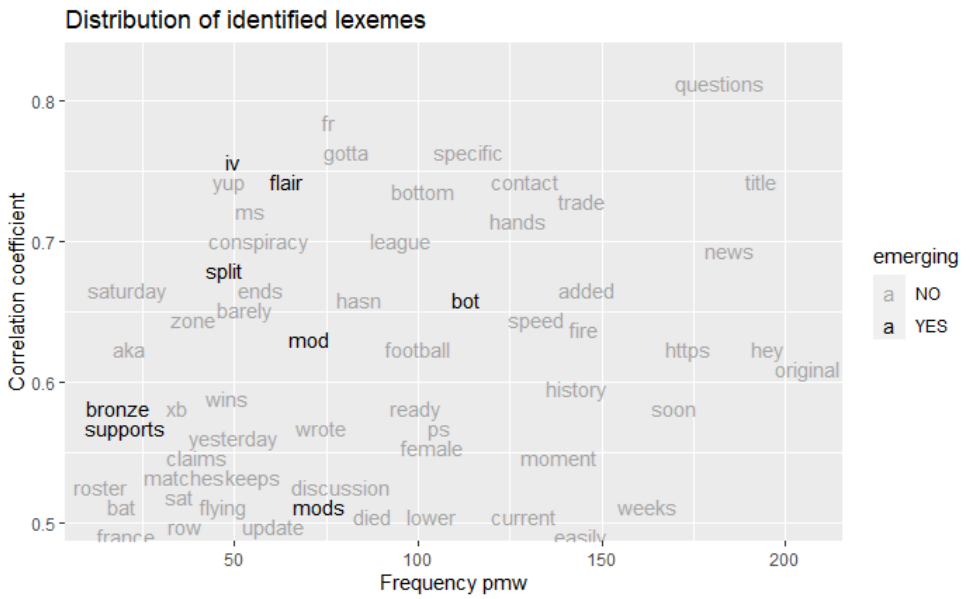
17 In their study, after gathering the data, Grieve *et al.* [2017: 103] then select the top 67,022 items for further analysis, which remained after choosing a minimum occurrence of 1,000 items as a cut-off point. Since the corpus in the present study was considerably smaller (13 million words compared to 8.9 billion words), a different threshold had to be chosen. As the smallest monthly data set of the year in question (April 2013) contains 6,960 word forms only, it was decided to pick the top 6,960 most frequent word forms from each monthly data set. This is of course a random line that could be drawn at any other number to produce a larger or smaller set of results. Drawing the line after the top 6,960 most frequent word forms, however, minimizes the number of empty slots in the calculations to follow.

18 In line with Grieve *et al.* [2017: 103-104], word forms were not lemmatized, and spelling variants were also treated as distinct items, since "alternative forms, including variant spellings, can often have different meanings or social distributions" [Grieve *et al.* 2017: 104]. Also similar to Grieve *et al.* [2017: 104], polysemous and homonymous words were not treated as separate items – an approach that can, of course, be questioned. Another question of interest at this point is what qualifies as a "word". Grieve *et al.* [2017: 103] define word forms as "a string of alphabetical characters plus hyphens, insensitive to case". The AntConc settings for the present study were therefore chosen to be case-insensitive and to treat any string of letters as a word.

19 It should also be noted that Grieve *et al.*’s [2017: 99] original intention was to analyse onomasiological change (“change in the way concepts are named, including the formation of new words”) only, but they later [2017: 122] discuss whether semasiological changes (established words adding a new meaning) should be included within the analysis (following the definition of onomasiological change by Geeraerts [2006: 38], which comprises semasiological change). Grieve *et al.* [2017: 122] rightly state that their methodology is unable to provide a full account of the semasiological changes taking place due to the higher initial frequency of the established word forms. Consequently, the word forms stemming from semasiological change that do appear in the results will not be disregarded, but no aspirations will be made to capture all semasiological changes taking place within the data set.

20 In accordance with the methodological steps outlined in Grieve *et al.* [2017: 103-107], the relative frequency per million words (hence pmw) was then calculated for all 6,960 word forms for every month of 2013. Using a Spearman rank correlation coefficient, the developments of the frequencies over the twelve months were then computed. On this basis, word forms starting off with a low relative frequency (less than 200 pmw in January) and increasing at a high rate during the year (correlation coefficient at least 0.5) were selected. This resulted in a list of 98 word forms, which are displayed in Figure 2 below, distributed by their correlation coefficient and their overall relative frequency in 2013.

Figure 2. Relation of relative frequency to correlation coefficient



21 The next step was to exclude proper nouns (like *France*) and established words (like *female*). Grieve *et al.* [2017: 107] cite “included in standard dictionaries” as their criterium for the latter. In accordance, the *OED* online (Oxford University Press [2020]) was used to check whether the usage of the established words on Reddit was covered by the dictionary entry or whether there were indications for semasiological change. After this sorting, a total of eight potential emerging lexical items remained.

2.3. Ethical considerations

22 Like every other study investigating natural language, studies on computer-mediated communication must properly consider the ethics of their procedure (Page *et al.* [2014: 58-59]). In the online environment, especially the privacy of the users is of relevance. Since in the present study only the raw text without any metadata (such as date of publication, subreddit, or username) is the object of analysis, anonymity is not an issue. However, even if the privacy of Reddit users can be guaranteed, the question of informed consent is likely to remain as unclear as in many other studies on social media data (Weller & Kinder-Kurlanda [2016: 169]). In their "Privacy Policy", Reddit [2020] states as follows:

> When you submit content [...] to the Services, any visitors to and users of our Services will be able to see that content, the username associated with the content, and the date and time you originally submitted the content. [...] Reddit also allows third parties to access public Reddit content via the Reddit API and via other similar technologies.

23 Based on this statement, one would expect Reddit users to be aware that the texts they produce might be accessed by other parties, including researchers.

3. Results

24 In the following sections, the emerging lexemes are shortly commented on, before their formal and semantic characteristics are described and illustrated in detail.

3.1. Identified emerging lexemes

25 The results can be viewed in Table 1 below, which displays the identified items, examples from the corpus, the respective correlation coefficient, an *OED* definition, and their specific use on Reddit that is not covered by the *OED*. Six of the eight items qualify for Grieve *et al.*'s [2017: 99] original intention of analysing onomasiological change, since an existing concept is assigned a new name or a new word is created for a new concept: *iv*, *mod*, *mods*, *bot* (*lane*), *split* (*push*), *bronze*. The other two appear to represent semasiological change, which includes established words adding a new meaning: *flair* and *supports*.

Table 1. Overview of potential emerging lexemes on Reddit

Word form	Coefficient	Example	OED-definition	Meaning in Reddit
bronze	0.59	the former *bronze* player / I'm *bronze*	(only for *bronze* as noun, *bronzed/bronzen* as adjective)	used as an adjective to describe the rank of players in the game "League of Legends" (other ranks are *silver*, *gold*, *platinum*, *diamond*)
iv	0.77	can offer a 5 *IV* eevee / give you 3 5 *IV* pokemon	short for *intravenously*	abbreviation for *individual values*, a strength score in the game "Pokémon"

supports	0.58	for most *supports* it's your job to / all kind of junglers/*supports*	the action or result of supporting, the action of supporting other armed forces, esp. by a second line of troops; organized assistance in a military, naval, or air force operation	denotes a certain role of players in multi-player games
flair	0.76	you need gray *flair* or better / *flair* up / please add *flair* to your post	power of 'scent', sagacious perceptiveness, instinctive discernment. Also: special aptitude or ability; liking, taste, enthusiasm	an optional picture or phrase that can be attached to a user's name within a specific subreddit
split (push / pusher)	0.69	*split pushing* capability / a great *split pusher*	a narrow break or opening made by splitting; a cleft, crack, rent, or chink; a fissure, A division formed by splitting	*(to) split push,* a tactic used in multi-player games in which one player splits away from the group
mod	0.64	vote for you as a *mod* / how we *mod* / release the *mod*	short for *modification*	*mod/mods* refer both to *modifications* made to computer games and to users assigned the role of a *moderator* within a subreddit
mods	0.52	the *mods* of this subreddit / install the *mods* accordingly	short for *modifications*	
bot (lane)	0.67	posted by a *bot* / defeats *bot lane* / we can push hard *bot*	short for *bot* (an automated program on a network, often having features that mimic human reasoning and decision-making; a program designed to respond or behave like a human); a software agent; short for *bottom*	robots imitating humans in online interactions, also a short version of *bottom* in the compound *bottom lane* (a route on the map of multi-player games)

26 It should also be noted that all identified items (or their homonyms) have an entry in the *Oxford English Dictionary*, which however does not contain the specific way in which the items are used on Reddit (alongside their traditional sense). Inspection of the concordance lines also suggests that *bot* and *split* are not single-word units, but part of a compound (*bot lane*, *split push/pusher/pushing*). The following section therefore looks at

the multi-word units as a whole and divides them into their components when necessary.

3.2. Formal characteristics

27 Now the formal characteristics of the identified word forms are described along the same lines used by Grieve *et al.* [2017: 107-120]. Looking at their word class, one can see that all forms except the adjective *bronze* are used as nouns, and that *mod*, *split push* and *flair* have an additional use as verbs (examples (01) to (03), emphasis added). This is not surprising, since nouns and verbs are open word classes which readily adopt new members.

> (01) how do I *flair* up? (Dec 2013)
> (02) I *mod* my reddit because I think its amusing to see the content people make on the subject I enjoy (Jan 2013)
> (03) If that doesn't work you can either *split push*, dive them on the turret, or [...] (Jun 2013)

28 Moving on to their word formation processes, *bot* (*lane*), *mod*, and *mods* appear to be the result of truncation. *iv* is the only example of an alphabetism, while the adjective *bronze* and the verbal use of *mod*/*mods* seem to result from conversion. *Bot lane* and *split push* are furthermore instances of compounding, whereas *supports* and *flair* are not formed by a word formation process, since an additional meaning has been added to an existing lexeme. In the case of *flair,* it appears that the new sense ("tag next to a username") originates from metaphoric extension of the old sense ("special aptitude or ability") [Traugott & Dasher 2002: 28]. The common characteristic in this case would be 'a feature that makes the possessor stand out'. For *supports*, metonymic extension (Traugott & Dasher [2002: 28]) from a general concept ("organised assistance in a military operation") to an individual associated with the concept ("role of players in computer games") seems to have taken place.

29 To analyse their recency, each of the word forms was searched for on Google Trends (Google Trends [2020]) and in the *Urban Dictionary* (*Urban Dictionary* [2020]), as it is done by Grieve *et al.* [2017: 110-111]. The results can be viewed in Table 2. One can see that some of the meanings (*bronze*, *supports*, *split push*, *bot lane*) appear to be so specific to the gaming community that they do not have an entry in the *Urban Dictionary*. The Google Trends tendencies, on the other hand, seem to be in line with the analysis so far, since some of the items feature an increase around the year 2013 (*bronze*, *split push*, *mod*, *mods*, *bot lane*), which could point to the emergence of the new usage. In total, the table shows that the new senses are either attested prior to their increasing frequency on Reddit or too specific to surface elsewhere.

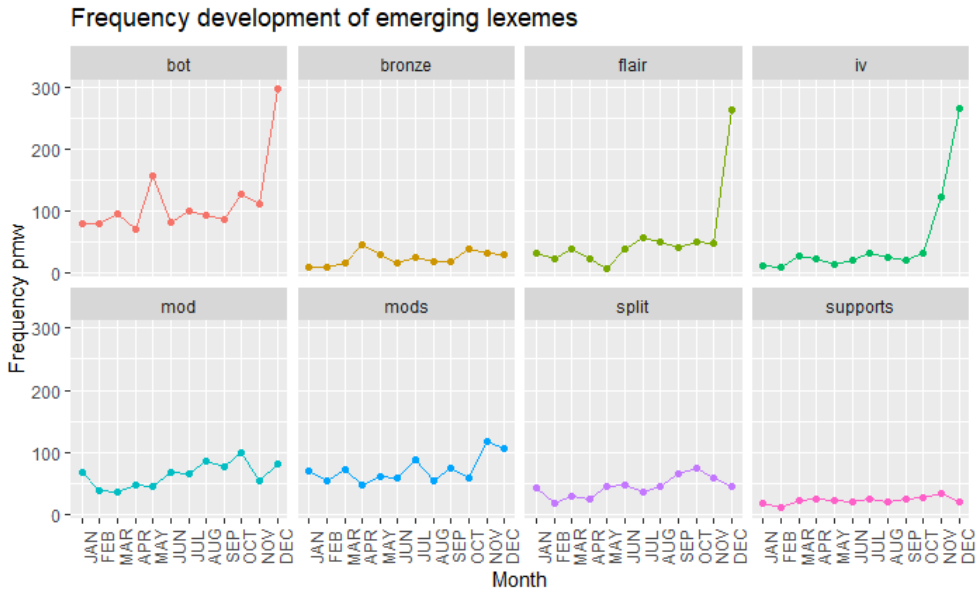
Table 2. The emerging lexemes on Google Trends and the *Urban Dictionary*

Word	*Urban dictionary*	Google Trends
bronze	-	slight increase since 2012
IV	2009	no noticeable increase/decrease

supports	-	no noticeable increase/decrease
flair	2008	increase since 2016
split (push)	-	no noticeable increase/decrease, *split push* increase after 2011
mod/mods	2003	increase from 2010 to 2014, then decrease
bot (lane)	(*robot*: 2002)	no noticeable increase/decrease, *bot lane* since 2011

30 Figure 3 now illustrates the change in relative frequency over the course of 2013 for each of the identified items. The trajectories show that the patterns are quite distinct from one another. While *iv*, *flair*, and *bot* only appear to increase at the end of the year (and potentially include considerable outliers in the data), the other items do not seem to follow any regular pattern. Some graphs can be explained by topicality trends due to major events during 2013: for example, the steep increase in the use of *iv* and *flair* is probably related to the release of a new edition of the "Pokémon" game in October of that year and the subsequent trading of the game's creatures on certain subreddits.

Figure 3. Frequency development of emerging lexemes in 2013



31 Summing up, one could say that most identified emerging lexemes belong to the expected word class of nouns and are formed by standard word formation processes (or add a meaning to an established word). Regarding their recent nature, some of the new uses appear to be attested before their rise in frequency. Looking at their trajectories during the year 2013 revealed different patterns of increase, which are likely to be related to their topicality, i.e. their relevance for certain events during the year. Based on these formal characteristics, it is now worth looking at the semantics in more detail.

3.3. Semantic characteristics

32 In this section, the semantics of the lexemes in question are considered. Readers might have noticed that the items appear to originate from two semantic domains only: online communication (*flair*, *mod*, *mods*) and gaming (*bot lane*, *mod*, *mods*, *split push*, *supports*, *iv*, *bronze*). This corresponds to other studies on Reddit, for example Kershaw *et al.* [2016] on language acceptance online, who also find many innovative word forms on Reddit and Twitter related to gaming language. While all of the lexemes theoretically qualify as "slang" (using the definition by Malmkjær [2010: 489]), most of them might be better described as "jargon" since they belong to the "specialist terminology" [Malmkjær 2010: 490] of the online gaming community or the Reddit community.

33 Worth investigating is also the "onomasiological competition with other lexical items" [Grieve *et al.* 2017: 117] for the analysed word forms. Some of them (*supports*, *split push*, *bronze*, *flair*) denote a specific concept and do not appear to have any obvious synonyms. Others (*bot lane*, *mod*, *mods*, *iv*) can be compared to their uncontracted parent forms. Figures 4 to 6 below illustrate the frequency developments of those word forms in 2013. It should be noted that the graphs show the development in frequency pmw for each item, instead of the percentages of each variable that Grieve *et al.* [2017: 117] calculate, due to the lexical ambiguity of the items, which makes it difficult to conceptualize them as variants of one lexical variable only.

Figure 4. Onomasiological competition of *bot* and *bottom*

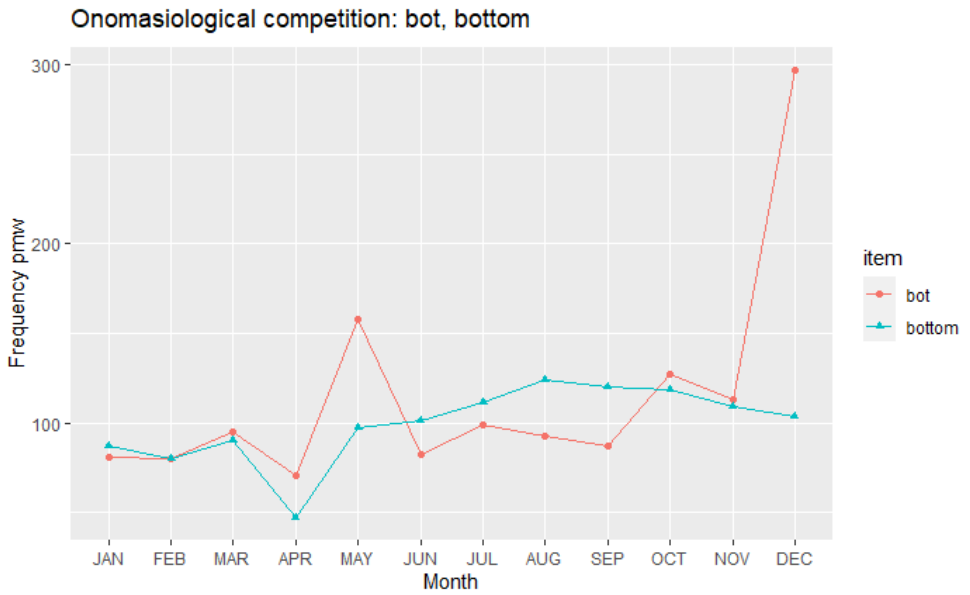


Figure 5. Onomasiological competition of *mod*, *mods*, *moderate*, and *moderators*

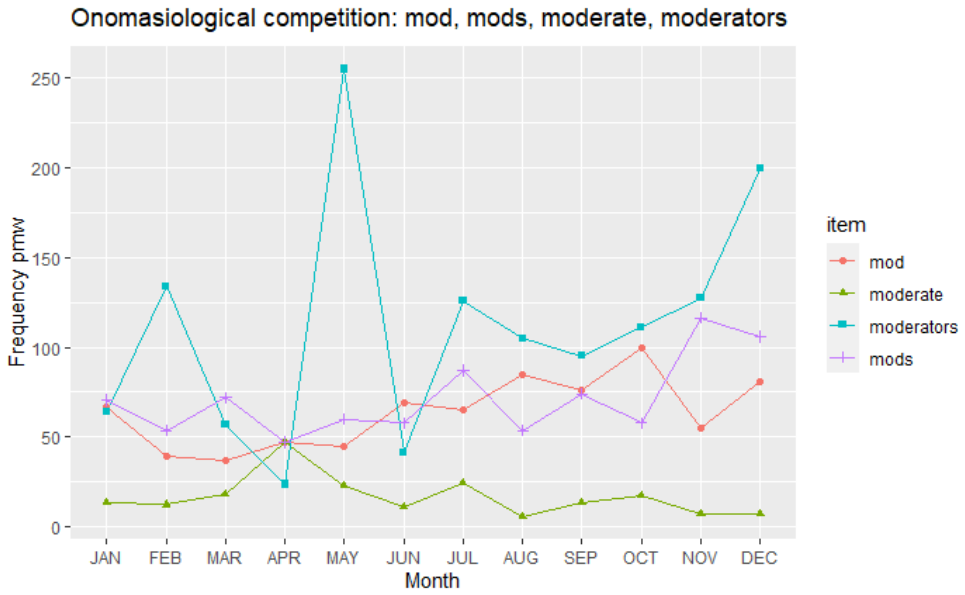
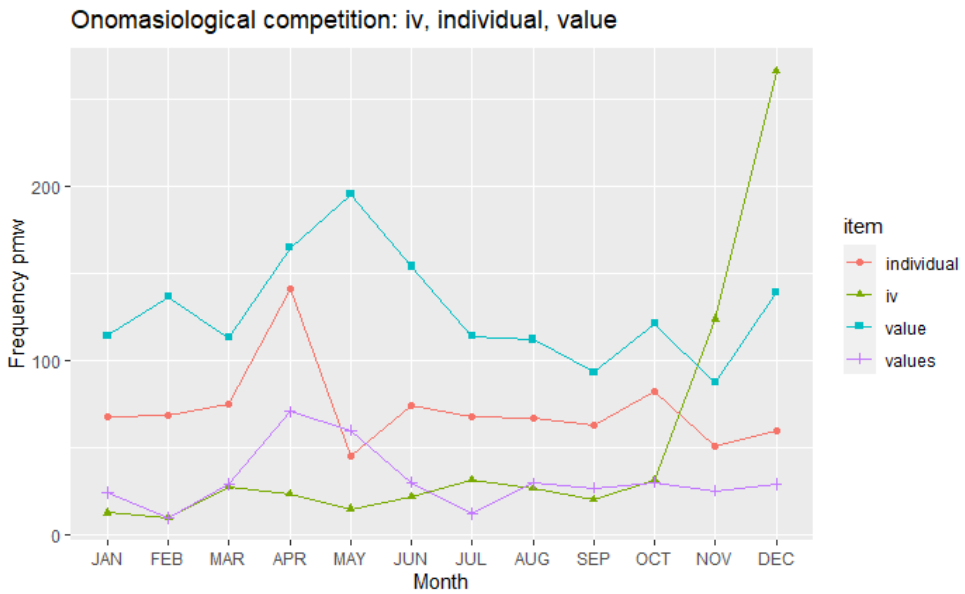


Figure 6. Onomasiological competition of *iv*, *individual*, *value*, and *values*



34 Considering *iv* and the words *individual* and *value/values*, the data suggests that there might also be a slight increase in the unreduced word forms as *iv* takes off. However, inspection of concordance lines for December does not reveal a single instance of *individual* and *value/values* being used together – this is therefore not an instance of competition. For *mod* and *mods*, as well as *bot* and *bottom*, the situation is more complex due to word class ambiguity and semantic ambiguity. While the substitution seems to be complete for the verb *moderate*, the plural form *moderators* is more frequent than its shortened equivalent *mods* (the singular form *moderator*, as well as *modify* and *modification*, are not part of the selected data set and cannot be compared).

35 Inspection of concordance lines does not indicate any semasiological change of the word forms resulting from onomasiological change during the year 2013. In conclusion, analysing the semantic characteristics of the emerging lexemes reveals that they stem from the semantic domains of online communication and online gaming, they can be described as either jargon or slang, and that their meaning does not appear to change during the time period investigated. Analysing their onomasiological competition shows that some lexemes are already institutionalised to a degree that they are no longer in competition with their uncontracted forms (*bot*, *iv*, *mod*), while others appear to be more interchangeable still (*mods*).

4. Discussion

36 The results presented in the previous section are now discussed in the light of findings by other researchers, most importantly the findings by Grieve *et al.* [2017]. The section focuses on the emerging word forms and their characteristics first, before commenting on the methodology proposed by Grieve *et al.* [2017].

4.1. Emerging lexemes

37 Looking at the formal characteristics first, the emerging lexemes identified in the Pushshift Reddit Dataset have several things in common with the items identified by Grieve *et al.* [2017] in their Twitter corpus. Both studies find that emerging lexemes belong to open word classes, especially nouns (Grieve *et al.* [2017: 108]). Their word formation processes are also similar in that mostly standard processes, particularly truncation, are used (Grieve *et al.* [2017: 108-109]). Grieve *et al.* [2017: 110] also find several instances of acronymization in their Twitter data, whereas only one (*iv*) is identified within the data set at hand. This might be due to the fact that Twitter has a lower character limit that motivates concise language, whereas Reddit allows for more characters. Grieve *et al.* [2017: 110-112] furthermore find that emerging lexemes might be attested and used infrequently for a long time before they increase in frequency. For those words that are attested outside of Reddit, this statement can be confirmed by the present analysis.

38 One noticeable difference, however, concerns the rate of the frequency change. Grieve *et al.* [2017: 112-117] claim that “the rate of change speeds up steadily over time” for most items, and that subsequently the frequencies either stabilize or decline again. They interpret these findings as proof for their hypothesis that language change online proceeds along the same s-shaped curves attested for ‘offline’ language change. The analysis at hand, however, reveals highly irregular patterns of frequency change. There are several possible explanations for this inconsistency: On the one hand, is it possible that the monthly structure of the Reddit data set is not able to adequately show the rate of change; a daily data set like the one used by Grieve *et al.* [2017] would be less prone to pick up such irregular patterns (see also the discussion of ‘granularity’, or “level of resolution” of the data by Gries [2012: 188]). On the other hand, it is possible that the rate of change is correlated to the idiosyncrasies of the online platform investigated, or that the theory does not apply (meaning that language change online does not follow s-shaped curves). At the moment, the structure of the data set seems the more plausible explanation.

39 Moving on to the semantic characteristics of the emerging lexemes, one can see that the semantic domains of the emerging lexemes on Twitter and Reddit are quite distinct. Grieve *et al.* [2017: 108] state that their items stem from the areas of “profanity and insult”, “recreational drug use”, “social media”, and “family and friends”. The word forms in the present analysis, on the other hand, belong to the areas of online gaming and online communication. This discrepancy can be explained by the special characteristics of the medium, its purpose and its users. While both platforms seem to be popular with younger, urban people (Eisenstein *et al.* [2014: 2], Duggan & Smith [2013]), Reddit is more topically organized and used as a discussion platform especially by the gaming community. This can be seen, for example, by looking at the most popular discussion topics on Reddit as presented by the “Pushshift” website (Baumgartner [2020]). At the time of writing, the most active subreddits include (among others): “r/gaming”, “r/leagueoflegends”, and “r/FortNiteBR”.

40 This divergence can also explain the language domains the emerging word forms belong to. While Grieve *et al.* [2017: 107-108] classify all their results as slang, this is only partially true for the items at hand. Most of them could be better categorized as technical jargon for online games and online communication. Similar is, however, that both studies reveal complex relationships between the emerging lexemes and their near synonyms (Grieve *et al.* [2017: 117-119]). Furthermore, the study at hand is unable to identify semasiological change (for the word forms resulting from onomasiological change) during the period investigated and Grieve *et al.* [2017: 119-120] only find one instance of this (*on fleek*).

41 In sum, the present analysis is able to confirm most of the findings by Grieve *et al.* [2017]. It is also found, however, that some characteristics of the emerging lexemes (semantic domain, language domain, and potentially also the rate of change) are closely connected to the individual character and properties of the online platform investigated. This context-dependence of lexical change was previously only attested for the offline environment. For example, Geeraerts *et al.* [2012: 128] emphasize how text type influences the emergence of the lexeme *anger*, and Hilpert [2012: 153-156] elaborates on the effect of genre on collostructional development. The results therefore highlight the danger of generalizing from one online platform to computer-mediated communication on the whole.

4.2. Methodology

42 This section now comments on the methodology described by Grieve *et al.* [2017] and employed in this study. The observations are presented roughly in the order of the corresponding methodological steps.

43 One aspect that Grieve *et al.* [2017: 104] arguably do not pay enough attention to is word class ambiguity, as well as polysemy and homonymy. If a word form belongs to several word classes or has several distinct meanings (for example *mod*) the frequency increase is distorted, since the different usages might show different frequency patterns. To solve the problem of word class ambiguity, it could be possible to run a parts-of-speech tagger over the data prior to the creation of the ranked wordlists. This might, however, not be as easy as it seems, since most taggers are likely to not accurately classify emerging lexemes (see also Liimatta [2016: 21]). By way of trial, the TagAnt programme (Anthony [2016]) was applied to one of the monthly sub-corpora to illustrate the

misclassification. Examples (04)-(05) show the result of this tagging, giving only the tags for the lexeme *mod*, which is classified as a noun in both comments, despite serving as a verb in the second utterance:

> (04) You reported me for "man hating" and had your mensrights loser mod_NN ban me from advice animals (June 2013)
> (05) You have to mod_NN the game (June 2013)

44 After the preparation of the list with potential emerging lexemes, Grieve *et al.* [2017: 107] winnow the items and remove all established words. As a criterion they name "words that are included in standard dictionaries". This appears to be a rather arbitrary selection criterion, since the resulting word forms will depend heavily on which dictionary is used and how up to date the entries are. It is, however, more systematic than the mere subjective judgement used by Stewart & Eisenstein [2018: 3]. In their study on language acceptance, Kershaw *et al.* [2016] employ a different method: for them, a word is classified as an innovation only if it has no search results in the British National Corpus. Using a linguistic corpus for comparison seems reasonable, since it reflects actual language usage instead of a lexicographer's perspective of language usage. The question which corpus to use depends on which online platform and which time period is being studied. For the analysis of Twitter and Reddit in 2013, the Corpus of Contemporary American English (Davies [2019]) appears to be a good choice, since it covers the year in question and contains written and spoken American English – the national variety both platforms are rooted in.

45 From the description by Grieve *et al.* [2017: 107] it furthermore remains unclear whether only the surface form was compared or whether the actual usage of a word form on Twitter was compared to the word senses listed in the dictionary. The present study shows how important it is to evaluate in detail the semantics within the corpus and the semantics attested elsewhere. Only that way can newly emerging meanings for established word forms be discovered as well. So far, there appears to be no reliable method absolving the researcher working with quantitative approaches from manual inspection of the concordance lines for the word forms in question (but see Sagi *et al.* [2012] for promising advancements towards the automatic detection of semasiological change).

46 As a further point, Grieve *et al.* [2017: 103] acknowledge that their methodology only allows for the detection of single-word units, which does not result in a comprehensive account of the lexical change taking place (as seen for *split push* and *bot lane*); Sang [2016: 8] encounters the same problem, as do many automatic neologism detection programs (Kerremans & Prokić [2018: 264]). Close inspection of concordance lines or using a collocates analysis tool can reveal emerging compounds or phrases as well. A further candidate for an emerging collocation is provided in examples (06)-(07), which was detected by having a closer look at the increasing frequency of the established lexeme *awkward*.

> (06) Insert Socially Awkward Penguin meme here. (Jan 2013)
> (07) I'm an awkward penguin in real life (Dec 2013)

47 To sum up, the methodology described by Grieve *et al.* [2017] is used effectively within this paper to detect onomasiological change and is expanded to incorporate instances of semasiological change as well. Some recommendations can nevertheless be made,

including the comparison with a corpus instead of a dictionary and the use of a tagging programme to solve the problem of word class ambiguity. On a more general note, this analysis emphasises the pivotal importance of close inspection of concordance lines in order to detect all different usages and potential compound forms of an item.

Conclusion

48 For convenience, the two research questions stated at the beginning are repeated below:
(i) Are the characteristics of emerging lexemes on Reddit similar to the characteristics of emerging lexemes on Twitter as identified by Grieve *et al.* [2017]?
Regarding the first question, the analysis shows that the emerging lexemes are overall similar concerning their formal and semantic characteristics. While most of the discrepancies can be explained by the individual character of the online platform used (which provides support for the context-dependence of lexical emergence), the question of the rate of change remains unanswered.

49 (ii) How applicable is the methodology outlined by Grieve *et al.* [2017] for the study of lexical emergence in a different online platform, Reddit?
Turning to the second question, the methodology proves to successfully identify newly emerging word forms and new meanings of established words. However, it needs to be mentioned that the precision of the method is relatively low, as a considerable number of established words are also identified as results as well (a problem also common in automatic neologism detection, see Kerremans & Prokić [2018: 251]). Some suggestions are made concerning the methodological steps; these concern the point of orientation for classifying a word form as “established”, the general importance of concordance line inspection, and an attempt at resolving the problem of word class disambiguation. This paper therefore tries not to invalidate the methodology but to make future researchers aware of the caveats and to suggest some possible amendments.

50 However, one also has to keep the limitations of the present study in mind. First of all, the Reddit data set is considerably smaller than that used by Grieve *et al.* [2017] and might include partial sampling errors, which could have led to slightly biased results. Furthermore, the different temporal resolutions (monthly instead of daily data sets) are likely to have statistical repercussions. These drawbacks notwithstanding, the present study is able to contribute to the state of knowledge in several ways: It tests and revises a relatively effective methodology for the comprehensive analysis of lexical emergence in the online environment with the help of large-scale corpora. It also illustrates the specific lexical characteristics of the online platform Reddit, especially in comparison to Twitter. More research on lexical emergence in the online environment is nevertheless needed. Further studies could widen the scope to other online platforms. An obvious next step would also be to extend the temporal limit of one year and see how a variation in the time frame affects the resulting lexemes (a first step in this direction has already been taken by Sang [2016] for Twitter).

BIBLIOGRAPHY

ALLAN Kathryn & ROBINSON Justyna (Eds.), 2012, *Current Methods in Historical Semantics,* Berlin: De Gruyter.

ANTHONY Laurence, 2016, *TagAnt (Version 1.2.0)*, Tokyo.

ANTHONY Laurence, 2018, *AntConc (Version 3.5.7)*, Tokyo.

BAKKEN Kristin, 2006, “Lexicalization”, *in* BROWN Keith (Ed.), *Encyclopedia of Language & Linguistics*, Elsevier, 106-108.

BAUMGARTNER Jason, 2020, *Pushshift*, https://pushshift.io/.

BAUMGARTNER Jason, ZANNETTOU Savvas, KEEGAN Brian, SQUIRE Megan & BLACKBURN Jeremy, 2020, “The Pushshift Reddit Dataset”, *in* ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE, *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, 830–839.

BITTORRENT INC, 2018, “µTorrent Web”, https://www.utorrent.com/intl/en/utweb-index.

BLYTHE Richard & CROFT William, 2012, “S-Curves and the Mechanisms of Propagation in Language Change”, *Language* 88(2), 269–304.

BRINTON Laurel & TRAUGOTT Elizabeth, 2005, *Lexicalization and Language Change*, Cambridge: Cambridge University Press.

BROWN Keith (Ed.), 2006, *Encyclopedia of Language & Linguistics*, Elsevier.

COLE Jeremy, GHAFURIAN Moojan & REITTER David, 2017, “Is Word Adaptation a Grassroots Process? An Analysis of Reddit Communities”, *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, Washington: Springer, 1–6.

DAVIES Mark, 2019, *Corpus of Contemporary American English. One billion words, 1990-2019*, https://www.english-corpora.org/coca/.

DEL TREDICI Marco & FERNÁNDEZ Raquel, 2017, “Semantic Variation in Online Communities of Practice”, *12th International Conference on Computational Semantics (IWCS)*, 1-13.

DEL TREDICI Marco & FERNÁNDEZ Raquel, 2018, “The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities”, *Proceedings of the 27th International Conference on Computational Linguistics*, 1591-1603.

DUGGAN Maeve & SMITH Aaron. *6% of Online Adults are Reddit Users. Young Men are Especially Likely to Visit the “Frontpage of the Internet”*, Washington: Pew Research Center.

EISENSTEIN Jacob, O’CONNOR Brendan, SMITH Noah & XING Eric, 2014, “Diffusion of Lexical Change in Social Media”, *PLoS ONE* 9(11), 1–13.

FISCHER Roswitha, 1998, *Lexical Change in Present-Day English; A Corpus-Based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*, Tübingen: Gunter Narr.

GEERAERTS Dirk, 2006, “Onomasiology and Lexical Variation”, *in* BROWN Keith (Ed.), *Encyclopedia of Language & Linguistics*, Elsevier, 37–40.

GEERAERTS Dirk, 2009, *Theories of Lexical Semantics*, Oxford: Oxford University Press.

GEERAERTS Dirk, GEVAERT Caroline & SPEELMAN Dirk, 2012, “How Anger Rose: Hypothesis Testing in Diachronic Semantics”, *in* ALLAN Kathryn & ROBINSON Justyna (Eds.), *Current Methods in Historical Semantics*, Berlin: De Gruyter, 109–131.

GOOGLE TRENDS, 2020, www.google.com/trends.

GRIES Stefan, 2012, “Commentary: Corpus-Based Methods”, *in* ALLAN Kathryn & ROBINSON Justyna (Eds.), *Current Methods in Historical Semantics*, Berlin: De Gruyter, 184–195.

GRIEVE Jack, 2018, “Natural Selection in the Modern English Lexicon”, *Proceedings of the 12th International Conference on the Evolution of Language (Evolang12)*, 153-157.

GRIEVE Jack, NINI Andrea & GUO Diansheng, 2017, “Analyzing Lexical Emergence in Modern American English Online”, *English Language and Linguistics* 21(1), 99–127.

HILPERT Martin, 2012, “Diachronic Collostructional Analysis. How to Use it and How to Deal with Confounding Factors”, *in* ALLAN Kathryn & ROBINSON Justyna (Eds.), *Current Methods in Historical Semantics*, Berlin: De Gruyter, 133–160.

KERREMANS Daphné & PROKIĆ Jelena, 2018, “Mining the Web for New Words: Semi-Automatic Neologism Identification with the NeoCrawler”, *Anglia* 136(2), 239–268.

KERSHAW Daniel, ROWE Matthew & STACEY Patrick, 2016, “Towards Modelling Language Innovation Acceptance in Online Social Networks”, *in* BENNETT Paul, JOSIFOVSKI Vanja, NEVILLE Jennifer & RADLINSKI Filip (Eds.), *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, New York: ACM Press, 553–562.

LIIMATTA Aatu, 2016, *Exploring Register Variation on Reddit: A Multi-dimensional study,* Helsinki: University of Helsinki.

MALMKJÆR Kirsten, 2010, *The Routledge Linguistics Encyclopedia*, New York: Routledge.

MEDVEDEV Alexey, LAMBIOTTE Renaud & DELVENNE Jean-Charles, 2017, “The Anatomy of Reddit: An Overview of Academic Research”, *Dynamics on and of Complex Networks*, 183–204.

NEVALAINEN Terttu & RAUMOLIN-BRUNBERG Helena, 2003, *Historical Sociolinguistics: Language Change in Tudor and Stuart England*, London: Pearson Education Limited.

OXFORD UNIVERSITY PRESS, 2020, *OED Online*, https://www.oed.com/.

PAGE Ruth, BARTON David, UNGER Johann & ZAPPAVIGNA Michele, 2014, *Researching Language and Social Media: A Student Guide*, Oxon: Routledge.

PAVLOV Igor, 2018, *7-Zip*, http://www.7-zip.de/.

R CORE TEAM, 2020, *R. A Language and Environment for Statistical Computing*, https://www.r-project.org/.

REDDIT INC, 2020, *Reddit Homepage*, https://www.redditinc.com/.

REDDIT INC, 2020, *Reddit Privacy Policy*, https://www.redditinc.com/policies/privacy-policy.

SAGI Eyal, KAUFMANN Stefan & CLARK Brady, 2012, “Tracing Semantic Change with Latent Semantic Analysis”, *in* ALLAN Kathryn & ROBINSON Justyna (Eds.), *Current Methods in Historical Semantics*, Berlin: De Gruyter, 161–183.

SANG Erik, 2016, “Finding Rising and Falling Words”, *in* HINRICHS Erhard, HINRICHS Marie & TRIPPEL Thorsten (Eds.), *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4HD)*, Osaka, 2–9.

SMITH Chris, 2016, “Tracking Semantic Change in fl - Monomorphemes in the Oxford English Dictionary”, *Journal of Historical Linguistics* 6(2), 165–200.

STEWART Ian & EISENSTEIN Jacob, 2018, “Making “fetch” Happen. The Influence of Social and Linguistic Context on Nonstandard Word Growth and Decline”, *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, 1-11.

TRAUGOTT Elizabeth & DASHER Richard, 2002, *Regularity in Semantic Change*, Cambridge: Cambridge University Press.

URBAN DICTIONARY, 2020, https://www.urbandictionary.com/define.php?term=.

WELLER Katrin & KINDER-KURLANDA Katharina, 2016, “A Manifesto for Data Sharing in Social Media Research”, *in* NEJDL Wolfgang, HALL Wendy, PARIGI Paolo & STAAB Steffen (Eds.), *Proceedings of the 8th ACM Conference on Web Science - WebSci ‘16*, New York: ACM Press, 166–172.

ABSTRACTS

The current advancements in the availability and size of electronic corpora, especially containing computer-mediated language, open up new possibilities for the study of change in the English lexicon [Allan & Robinson 2012: 4]. In line with these developments, Grieve *et al.* [2017] present a methodology for finding “emerging lexemes” and apply it to a corpus of American Twitter data from 2013 to 2014. Their methodology entails searching for word forms that start off with a low overall frequency and that feature a high correlation coefficient with their rank in the time series over the whole year [Grieve *et al.* 2017: 103-105]. Working with a one-year section of the Pushshift Reddit Dataset (Baumgartner *et al.* [2020]), this study applies the methodology proposed to a different online forum, Reddit.

The present paper therefore has two aims: to test the methodology proposed by Grieve *et al.* [2017] and to investigate recent lexical emergence on the platform Reddit. This also allows for a comparison between the two platforms Reddit and Twitter to provide further insights into the context-dependence of lexical emergence in the online environment. Furthermore, the trial and refinement of the methodology for discovering emerging lexemes holds valuable insights for scholars looking to use this procedure in the future.

Applying the methodology to the Pushshift Reddit Dataset yields a total of eight emerging lexemes; six resulting primarily from onomasiological change, while two appear to be the outcome of semasiological change. The formal characteristics of the emerging lexemes (word class, word formation process) are overall very similar to the features identified by Grieve *et al.* [2017: 108-109], whereas their trajectories over the time period investigated vary noticeably and do not follow the s-shaped curves that are commonly proposed [e.g. Blythe & Croft 2012] and that are also attested by Grieve *et al.* [2017: 116]. Concerning the semantic criteria, the semantic domains of the identified lexemes differ considerably from the results by Grieve *et al.* [2017: 107-108], which can also be explained by the different profiles of the two platforms and their users.

Several caveats could be identified for the application of the methodology by Grieve *et al.* [2017]: first of all, word class ambiguity is likely to distort the frequencies obtained. Secondly, words being attested in a representative corpus was proposed as a more realistic criterion for classifying a word as ‘established’ compared to its inclusion in standard dictionaries. A third problem is that the methodology only allows for the detection of single-word units, which is not an accurate representation of the changes taking place, as several of the emerging lexemes appear to be part of compounds.

Les progrès actuels concernant la disponibilité et la taille des corpus électroniques, particulièrement contenant du langage virtuel, créent de nouvelles possibilités de recherche dans le milieu des changements sémantiques du lexique anglais [Allan & Robinson 2012 : 4]. En adéquation avec ces développements, Grieve *et al.* [2017] présentent une méthodologie destinée à trouver de « nouveaux lexèmes » et l'appliquent à un corpus de données concernant Twitter aux États-Unis de 2013 à 2014. Leur méthodologie implique une recherche de mots débutant avec une fréquence globale basse et présentant un coefficient de corrélation élevé avec leur rang dans les séries chronologiques sur l'ensemble de l'année. [Grieve *et al.* 2017 : 103-105]. Se concentrant sur une période d'un an du Pushshift Reddit Dataset (Baumgartner *et al.* [2020]), cette étude applique la méthodologie proposée à un forum en ligne différent : Reddit.

Par conséquent, cet article a deux objectifs : tout d'abord, tester la méthodologie proposée par Grieve *et al.* [2017], puis étudier les émergences lexicales récentes sur la plateforme Reddit. Cela permettra également une comparaison entre les deux plateformes Reddit et Twitter dans le but de fournir un éclairage complémentaire sur la dépendance au contexte de l'émergence lexicale dans un environnement virtuel. De plus, la mise à l'épreuve ainsi que le perfectionnement de la méthodologie permettant de découvrir de nouveaux lexèmes fourniront des informations précieuses pour les spécialistes souhaitant par la suite utiliser cette procédure.

Appliquer la méthodologie au Pushshift Reddit Dataset permet d'observer un total de huit nouveaux lexèmes ; six résultant principalement d'un changement onomasiologique, ainsi que deux apparaissant comme le résultat d'un changement sémasiologique. Les caractéristiques des nouveaux lexèmes (catégorie grammaticale, processus de formation lexicale) sont en général très similaires à celles identifiées par Grieve *et al.* [2017 : 108-109], alors que leurs trajectoires durant la période étudiée varient radicalement et ne suivent pas les courbes en S généralement proposées [par ex. Blythe & Croft 2012] et également attestées par Grieve *et al.* [2017 : 116]. Concernant les critères sémantiques, les domaines sémantiques correspondant aux lexèmes identifiés diffèrent considérablement des résultats obtenus par Grieve *et al.* [2017 : 107-108], ce qui peut également être expliqué par les profils différents des deux plateformes ainsi que par la différence de leurs utilisateurs respectifs.

De nombreuses limites ont pu être identifiées concernant l'application de la méthodologie de Grieve *et al.* [2017] : tout d'abord, une possible ambiguïté liée à la catégorie grammaticale est susceptible de fausser les fréquences obtenues. Dans un second temps, la classification des mots en tant que mots ‘établis’ a été basée sur la présence attestée de ces mots dans un corpus représentatif plutôt que sur leur présence dans les dictionnaires standards. Dans un troisième temps, la méthodologie permet seulement la détection de lemmes simples, ce qui ne représente pas le changement actuel de manière exacte, puisque de nombreux nouveaux lexèmes semblent faire partie de mots composés.

INDEX

Mots-clés: émergence lexicale, Reddit, Twitter, communication virtuelle
Keywords: lexical emergence, Reddit, Twitter, computer-mediated communication

AUTHOR

HANNA MAHLER

Otto-Friedrich-Universität Bamberg
hanna.mahler@uni-bamberg.de